2015

New and not so new methods for assessing oral communication

Gary Ockey
Iowa State University, gockey@iastate.edu

Zhi Li
Iowa State University

New and not so new methods for assessing oral communication

Abstract
The assessment of oral communication has continued to evolve over the past few decades. The construct being assessed has broadened to include interactional competence, and technology has played a role in the types of tasks that are currently popular. In this paper, we discuss the factors that affect the process of oral communication assessment, current conceptualizations of the construct to be assessed, and five tasks that are used to assess this construct. These tasks include oral proficiency interviews, paired/group oral discussion tasks, simulated tasks, integrated oral communication tasks, and elicited imitation tasks. We evaluate these tasks based on current conceptualizations of the construct of oral communication, and conclude that they do not assess a broad construct of oral communication equally. Based on our evaluation, we advise test developers to consider the aspects of oral communication that they aim to include or exclude in their assessment when they select one of these task types.

Keywords
oral communication, speaking, assessment, methodology

Disciplines
Bilingual, Multilingual, and Multicultural Education | Curriculum and Instruction | Educational Assessment, Evaluation, and Research | Educational Methods

Comments
This is an article from *Language Value* 7 (2015): 1, doi:10.6035/LanguageV.2015.7.2. Posted with permission.

Language Value November 2015, Volume 7, Number 1 pp. 1-21
http://www.e-revistes.uji.es/languagevalue

New and not so new methods for assessing oral communication

Gary J. Ockey
gockey@iastate.edu
Zhi Li
zlisu2010@gmail.com
Iowa State University, USA

ABSTRACT

The assessment of oral communication has continued to evolve over the past few decades. The construct being assessed has broadened to include interactional competence, and technology has played a role in the types of tasks that are currently popular. In this paper, we discuss the factors that affect the process of oral communication assessment, current conceptualizations of the construct to be assessed, and five tasks that are used to assess this construct. These tasks include oral proficiency interviews, paired/group oral discussion tasks, simulated tasks, integrated oral communication tasks, and elicited imitation tasks. We evaluate these tasks based on current conceptualizations of the construct of oral communication, and conclude that they do not assess a broad construct of oral communication equally. Based on our evaluation, we advise test developers to consider the aspects of oral communication that they aim to include or exclude in their assessment when they select one of these task types.

Keywords*: oral communication, speaking, assessment, methodology*

Introduction

Practice and research in assessing oral communication is regarded as the "youngest sub-field in language testing" (Fulcher 2003, p. 13). The testing process, the construct to be measured, the tasks used to measure the construct, and the technology used to aid in the process all continue to evolve as the field matures. These developments have helped to minimize some of the challenges that are faced in the assessment of oral communication. In this paper, we discuss the current state of the assessment of second language oral communication in the light of some of these developments. We begin by briefly outlining the oral communication assessment process. Our aim with this section is to provide an indication of some of the factors to be considered when assessing oral communication. Next, we provide an oral communication construct, which is in line with current conceptions in the field. The greater part of our paper is provided in the next section, which describes the tasks that are currently being used to aid in assessing

DOI: http://dx.doi.org/10.6035/LanguageV.2015.7.2

oral communication. Along with the description of these tasks, we analyze them based on the degree to which they can be used to effectively measure our construct of oral communication, given the factors presented in an oral communication assessment process.

I. THE ORAL COMMUNICATION ASSESSMENT PROCESS

A number of factors contribute to the score that a test taker receives on a test designed to assess his or her ability to communicate orally. Figure 1 provides a graphic display of how some of these factors affect scores.

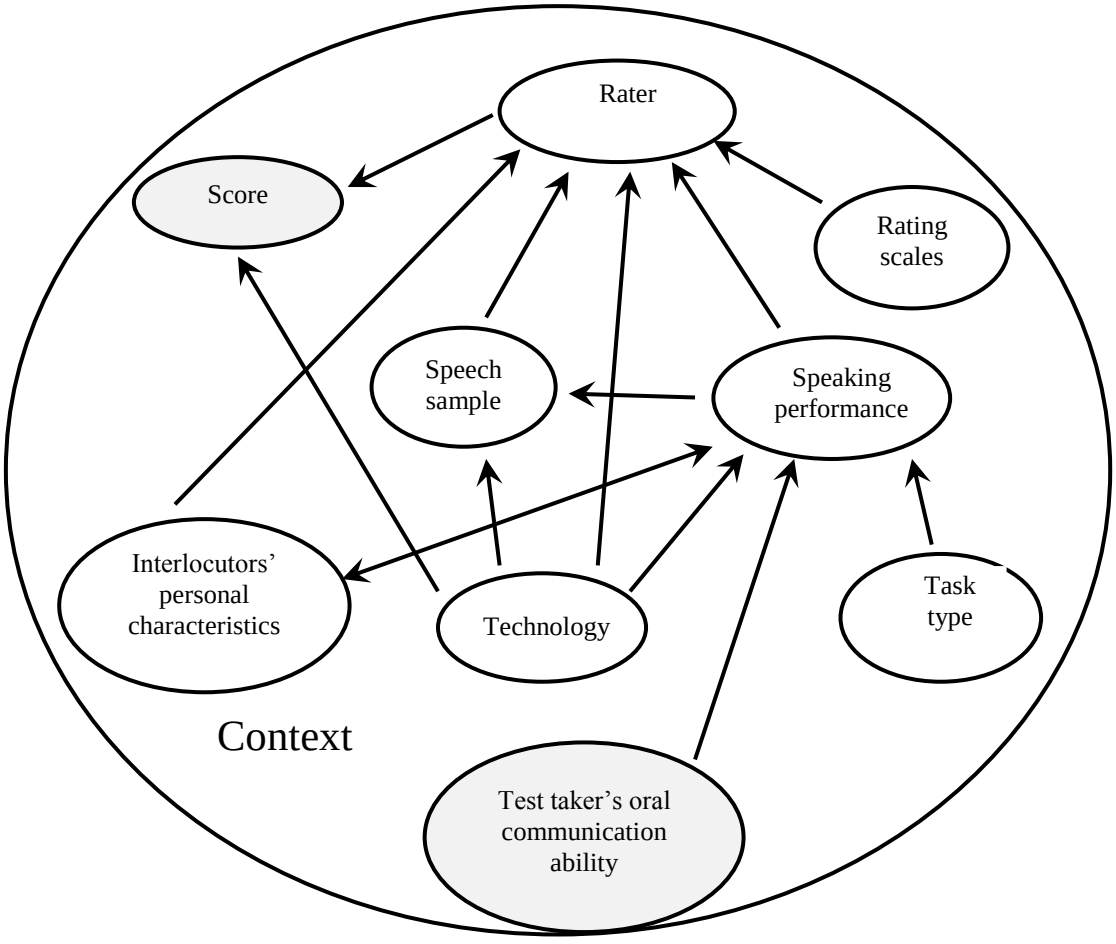

Figure 1 Model of assessment of oral communication

The conceptualization builds on the earlier models of Kenyon (1992), McNamara (1996), Skehan (1998), Bachman (2001), and Ockey (2009). The model focuses on factors that have an impact during the administration of the test. Other factors, such as the impact of the score on instruction, are not explicitly identified in the model, but are considered to be part of the testing context. In the figure, the test taker's oral communication ability is depicted by an oval at the bottom of the large circle. The aim of the assessment is to measure this ability. The score that is assigned to the test taker based on the assessment is shown in the upper left part of the large circle. This score is used to indicate the test taker's oral communication ability. As can be seen in the figure, task type, other interlocutors' personal characteristics, technology used for the assessment, the actual speaking performance and resulting speech sample, rating scales, and raters can all have an impact on scores during a test administration. These factors, coupled with the context (e.g., stakes, consequences, sociopolitical situation, and cultural expectations of stakeholders) of the assessment, may all be sources of construct-irrelevant variance in an oral communication assessment.

A number of factors have influence on test scores. Task types (as well as the specific prompts used for a particular task type) can affect a test taker's speaking performance. Familiarity with a task may be an advantage for test takers. For instance, some test takers may do better with a group or paired discussion because they are used to talking to others in a group setting. The personal characteristics of the other interlocutors involved in the assessment can influence speaking performance and can have an effect on how raters evaluate a test taker's ability. For example, in a one-on-one interview, an interviewer can affect the scores by being more or less supportive during the interview, and in a group/paired oral test, the members of the test taker's group might be very assertive, thus having an impact on the test taker's speaking performance. The rater may also compare the performance of the group members in a group/paired discussion, thereby making it possible for the abilities of the other members of one's group to have a direct effect on a test taker's score through the rater. Technology can also affect a test taker's speaking performance. For instance, a test delivered over the telephone or the internet may be interrupted by a slow or unclear connection, making it difficult for the test taker to understand the prompt. Technology can also affect the speech sample if, for example, the recording device does not function effectively. Technology might also

influence scores through the rater, who, for instance, may not be able to use the data entry procedures effectively. Finally, technology can affect a test score directly if the scoring system does not work properly.

Rating scales can also have an influence on scores. Since they are designed to measure the construct that the tester aims to assess, they therefore play a crucial role in linking the scores to the construct that the test is designed to measure. To be effective, rating scales must clearly reflect the construct and be easily interpretable. Raters can be human, computer (automated scoring), or a combination of both. Human raters and/or computer automated scoring engines play a key role in the oral communication assessment process, and can affect scores in several ways, depending on how they are trained or programmed to interpret the rating scales and evaluate elicited speech samples. All of these factors work in a given assessment context, which also influences the scores assigned to test takers. There are many contextual factors, including physical features of the setting, such as the temperature of the room and level of external noise, and psychological factors, such as the test taker's anxiety and motivation. Given all of these factors in an oral communication assessment process, it is crucial that test designers define the oral communication ability that they aim to assess as clearly as possible, and then consider each of these factors to best ensure a valid assessment process.

II. THE CONSTRUCT OF ORAL COMMUNICATION

The definition of the construct spells out the key components or essential aspects of the ability test developers wish to measure. In the context of assessing speaking, Fulcher (2003: 23) defined speaking ability as "the verbal use of language to communicate with others". Fulcher's (2003) definition is in line with other more recent definitions, such as that of Jamieson, Eignor, Grabe, and Kunnan (2008: 74), who defined speaking ability as "the use of oral language to interact directly and immediately with others… with the purpose of engaging in, acquiring, transmitting, and demonstrating knowledge". These broad definitions of oral proficiency suggest that this ability includes: 1) interactional competence; 2) appropriate use of phonology; 3) appropriate and accurate use of vocabulary and grammar; and 4) appropriate fluency.

Interactional competence can be viewed as an individual's underlying ability to actively structure appropriate speech in response to incoming stimuli, such as information from another speaker, in real time. That is, interactional competence can be considered as the individual attributes that test takers need to engage in real-time interactive communication, which may not be necessary to engage in non-interactive communication. More specifically, interactional competence entails the ability to take turns, open and close gambits, respond to others, and negotiate and develop topics with appropriate pragmatic use in a given context. Research suggests that interactional competence is not adequately assessed with description or prepared oral presentation tasks (Ockey, Koyama, Setoguchi, and Sun 2015). Other research indicates that test takers prefer real-time tasks in which they actively co-construct meaning with other interlocutors. They also feel that such tasks are better indicators of their oral communication ability in the second language (Brooks and Swain 2015).

Appropriate use of phonology relates to the effective use of both segmental and prosodic aspects of language. At the segmental level, pronunciation refers to the ability to articulate words and create the physical sounds that endow a word with a meaning. Prosodic aspects of phonology include stress, increased volume on a syllable, intonation, voice movement, and pitch (Fulcher 2003). A major conundrum in assessing second language oral communication relates to how to assess accent, an important aspect of phonology. Strength of accent has been defined as, "the degree to which (the accent) is judged to be different than the local variety, and how it is perceived to impact the comprehension of users of the local variety." This definition emerged partly as a result of research which has shown that high-stakes assessments that are rated by local raters who are familiar with the speakers' first language can assign much more lenient ratings than raters who are not familiar with the local first language (Carey, Mannel and Dunn 2011). While some argue for more acceptance of various accents when assessing oral communication (Abeywickrama 2013; Smith and Bisazza 1982), others note the importance of accent in oral communication, and argue that to be fair to test takers, oral communication assessments should carefully consider the accent of the input (Elder and Harding 2008; Ockey and French 2014) and judge the strength of the test takers' accent as a part of their oral communication ability.

Appropriate and accurate use of vocabulary and grammar refer to vocabulary breadth, how many words are known; vocabulary depth, how well and effectively the words are known and can be used (Nation 1990); grammatical breadth, how many grammar structures are known and can be used; and grammatical depth, how accurately and effectively these grammatical structures can be used. Grammar and vocabulary have been treated as separate constructs, but research suggests that human raters do not assign distinct scores for vocabulary and grammar in oral communication assessments (Batty 2006; Hunston, Francis and Manning 1997). Given the strong relationship between scores on vocabulary and grammar, it can be argued that they should be treated as one sub-ability of oral communication.

As one of the four components in the construct of oral communication, fluency, which refers to naturalness of rate of speech, pausing, and repetition has attracted a lot of attention. The temporal aspects of fluency include various measures of quantity, rate, pausing, and language repair (Bosker, Pinget, Guene, Sanders and de Jong 2012; Ginther, Dimova and Yang 2010). Research indicates that the temporal measures of fluency are important components of oral communication. Sato (2014) labeled fluency within interactions as interactional oral fluency and argued that fluency is a 'perceived phenomenon', in which temporal aspects of fluency are interwoven with interactional features.

Having laid out the process of assessing oral communication and the factors that affect it, along with the construct of this ability, we now turn to the types of task that have been used to assess oral communication. We describe each task and provide an analysis of the extent to which it assesses the construct of oral communication that we have provided, given the process and accompanying factors of assessing this ability.

III. TASK FORMATS IN TESTING ORAL COMMUNICATION SKILLS

Many of the types of tasks that are currently popular in testing oral communication skills are described by Harris (1969), who introduced three main types of "oral production tests" used in the 1960s, namely: 1) scored interviews, 2) highly structured speech samples, and 3) paper-and-pencil tests of pronunciation. The first type of task requires one or more trained interviewers/assessors to engage in conversations with test

takers and rate their performance based on established scales. Scored interviews and their variants are still widely used as one of the dominant tasks in assessing oral communication ability. The second type of tasks relies on pre-set stimuli and does not involve interlocutors. Typical examples in this category include sentence repetition and reading a passage aloud. The tasks in the family of highly structured speech samples have gone through ups and downs in the past decades but have seen a certain degree of revival in recent years, thanks at least in part, to the emergence of automated speech rating systems. The last type of tasks described in Harris (1969) requires written responses about finding rhyme words, identifying word stress and phrase stress. This type of paper-and-pencil tests of oral production has mostly disappeared from language testing, probably as a result of the popularity of the communicative language teaching paradigm.

We now discuss five popular types of task used to assess oral communication. The task types are oral proficiency interviews, paired/group oral discussion tasks, simulated tasks, integrated oral communication tasks, and elicited imitation tasks. The first three task types are variants of scored interviews, while the last two are variants of the highly structured samples discussed in Harris (1969).

III.1 Oral proficiency interviews

Oral proficiency interviews are one of the most commonly used task formats for assessing oral communication. A typical oral proficiency interview task requires a test taker to respond to questions on different topics posed by an interviewer, who usually chooses the topics, initiates the conversations, and sometimes rates the speech samples elicited from the test taker in the test. In these tasks, test takers are expected to respond to questions but they usually have limited opportunities to demonstrate their ability to negotiate meaning, open and close gambits, or elicit opinions from the interviewer.

One example of this task format is the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI), which was developed in the 1980s and has been widely used as a "standardized procedure for the global assessment of functional speaking ability" in a number of foreign languages (ACTFL 2012; Liskin-Gasparro 2003). The task format used in the ACTFL OPI is a face-to-face or telephonic

interaction between a test taker and a certified interviewer on a series of personalized questions. In the OPI tests, an interviewer initiates and leads the conversations by choosing topics of a variety of natures, including personal, general, and abstract. In this testing context, the interviewer, who is not familiar with the test taker, controls the topics, asks the questions, and generally directs the course of the conversation. The ACTFL OPI test includes four mandatory phases, namely, warm-up, level checks, probes, and wind-down. In the warm-up step, the interviewer asks the test taker simple questions and establishes rapport. At the level checks step, the interviewer asks the test taker a number of questions with the aim of deciding the proficiency floor, or the proficiency level that the test taker can successfully demonstrate. In the probes step, the interviewer asks questions with a higher level of proficiency than the level expected by the test taker in order to determine the proficiency ceiling, or the highest possible proficiency level that the test taker can sustain.

There are two technology-mediated variants of the ACTFL OPI: telephonic OPI and the Internet-delivered version of OPI or OPIc. In the telephonic OPI, test takers call the testing center and take the test via phone, rather than face-to-face as in regular OPI test. Instead of involving a human interlocutor, the OPIc uses an avatar, through which the one-on-one interview model is simulated.

The speech samples elicited from the real time (face-to-face and telephonic) OPI-styled tasks make it possible to assess a number of aspects of oral communication, such as global tasks and functions, context and content, grammatical accuracy, and text type. Specifically, the ACTFL rating rubric contains detailed descriptions regarding test taker's performance in terms of fluency, pronunciation, grammar, and vocabulary. These aspects of oral communication closely resemble three of the four aspects of oral communication discussed in the construct section of the paper. However, OPI tasks are likely to fall short in assessing many aspects of interactional competence (Johnson and Tyler 1998; van Lier 1989). One of the major concerns is that the speech samples elicited from the OPI tasks do not exhibit key features observed in natural conversations, such as “reactive and mutual contingency” (van Lier 1989: 501), which refers to the spontaneous and interactive sequence of speech between two speakers. This may be related to the unequal power relationship between test takers and the interviewer (Johnson and Tyler 1998). The interviewer asks the questions, while the test takers can

only respond to what is asked. In addition, the interviewer's behaviors and degree of involvement can contribute to the "asymmetrical and pseudosocial" nature of the OPI interactions. That is, the discourse that the test taker produces can be affected by the personal characteristics of the interviewer. For example, the interviewer may be much less friendly than other interviewers, which can lead to an unrepresentative sample of discourse from the test taker (Brown 2003). The technology-mediated variants of the OPI-type tasks have similar limitations, but they probably assess even less of the interactional competence aspect of the oral communication construct, given that test takers cannot ask for clarification of a question or interact with the interviewer at all.

III.2 Paired and group oral discussions

Paired and group oral discussions can address some of the limitations of the OPI-type tasks. In this format, pairs or small groups of students have a discussion with each other. A trained interlocutor acts as a moderator and may or may not participate in certain aspects of the task. Test takers can be paired or grouped as equal status speakers based on different criteria, for example, proficiency level or interpersonal relationship. One example task is the group discussion in the College English Test – Spoken English Test (CET-SET), which uses a computer program to group three to four test takers and requires them to sustain a 4.5-minute face-to-face discussion on a given topic (He and Dai 2006).

The potential of group oral tasks in assessing oral communication was recognized in the 1980s and the last few decades have witnessed more implementations of this type of tasks in both high-stakes contexts such as the Cambridge Main Suite Examinations in the UK, including the First Certificate in English (FCE) and the Cambridge Certificate in Advanced English (CAE), CET-SET in China, and the speaking section of a provincial exit exam in Canada (Turner 2009), as well as various local English placement tests, such as the Kanda Assessment of Communicative English in Japan (**Ockey, Koyama, Setoguchi and Sun** 2015) and a placement test at Michigan State University in the USA (Winke 2013).

Currently, paired and group oral discussion tasks are mainly carried out in a face-to-face manner, thus requiring the physical presence of each participating test taker. However,

such tasks could be completed via synchronous voice-based computer-mediated communication (CMC), as has been done to aid in English teaching (Alastuey 2011; Lin 2014). For example, video-conference techniques, such as Adobe Connect and Skype, could be used as a testing platform to connect test takers who are not in a face-to-face context. In addition, computer technology could be used to group test takers based on pre-established criteria such as English proficiency level, personality traits, and topic familiarity.

Since paired and group oral tasks are designed to elicit interaction among test takers, accordingly, the rating rubric for paired and group oral tasks generally includes the sub-construct of interactional competence and can therefore accommodate a broader coverage of the oral communication construct than OPI-type tasks. For example, in the group oral placement test described in Bonk and Ockey (2003), test takers' performances are rated on pronunciation, fluency, grammar, vocabulary/content, and communicative skills/strategies, the latter being essentially another name for interactional competence. In the group discussion task of the CET-SET, the evaluative criteria include: 1) accuracy in pronunciation, stress/intonation, and use of grammar and vocabulary, 2) range of vocabulary and grammatical structures, 3) size (percentage) of contribution to group discussion, 4) discourse management, 5) flexibility in dealing with different situations and topics, and 6) appropriateness in the use of linguistic resources (Zheng and Cheng, 2008). These aspects of oral communication fit quite closely with the four aspects of oral communication described in the construct section above, thus suggesting that group and paired tasks aim to assess all four components of the construct. Empirically, it has been found that peer-to-peer discussion provides test takers with a better opportunity to demonstrate their ability to engage in complex interaction, compared with test taker-to-interviewer interaction, as is the case with OPIs (Brooks 2009). In this sense, paired and group oral tasks can tap into a fuller range of oral communication abilities than OPI-type tasks.

Given the complex interaction patterns exhibited in paired and group oral tasks, this task type has attracted much attention. With regard to the effects of interlocutor traits, research suggests that the test takers' familiarity with other test takers (O'Sullivan, 2002), as well as personality (level of extraversion), English proficiency level, and the number of participants, may influence test takers' performance in group discussion

tasks, as shown in a study on Japanese secondary school students conducted by Nakatsuhara (2011). On the other hand, Ockey, Koyama, and Setoguchi (2013) investigated the effect of interlocutor familiarity on test takers' performance in a Japanese university. A comparison of the scores of the two groups of test takers, namely a class-familiar group and a class-unfamiliar group, showed that interlocutor familiarity did not exert a significant influence on four rating categories (pronunciation, fluency, lexis and grammar, and communication skills), suggesting that at least for some contexts, interlocutor familiarity may not have a significant impact on scores elicited from the group oral. The effects of prompts in oral discussion tasks are reported by Leaper and Riazi (2014), who compared the turn-taking features, syntactical complexity features, accuracy, and fluency in the test taker's discourse elicited with four prompts. It was found that the prompts that allowed for an account or extension of personal experiences tended to elicit longer and more complex turns, whereas the prompts with factual content yielded shorter and less complex turns.

To sum up, the group and paired oral tasks have the advantage of providing test takers with the opportunity to demonstrate their interactional competence. This opportunity seems to stem from the rather loose controls placed on the task. That is, test takers seem to be able to demonstrate their interactional competence because the task affords them a fair number of opportunities to collaborate with others of equal status. On the other hand, because of this loose control, the task is susceptible to a number of factors, such as the personalities of other test takers with whom they are grouped, which can affect their test scores.

III.3 Simulated tasks

Simulated tasks are commonly used to assess oral communication in the context of English for specific purposes (ESP). An example of this type of task is role-play tasks, which require a test taker to assume a particular role in a simulated task context, for example, a meeting with a professor during office hours. Another example of a simulated task is the teaching tasks used in assessing the oral communication ability of prospective international teaching assistants (ITAs) in English-speaking universities. The Taped Evaluation of Assistants' Classroom Handling (TEACH), originally

developed at Iowa State University in 1985, is a performance test with simulated tasks for ITAs (Papajohn 1999; Plakans and Abraham, 1990). The TEACH test consists of three phases including a 2-minute preparation, a 5-minute lecture, and a 3-minute question-answering activity. As introduced in Papajohn (1999), some undergraduates are invited to the testing room to form a 'mock class', and ask questions to the ITAs in the TEACH test. In these simulated tasks, test takers select topics in their own field as the teaching content and present the lecture to mock students as well as the assessors.

Another variant of simulated tasks attempts to assess pragmatic competence through computerized discourse completion tasks (DCT) with a video prompt (Sydorenko, Maynard and Guntly 2014). In this task, test takers are presented with video prompts which describe situations requiring them to make appropriate requests. The test takers respond to the prompts orally and then computer technology is used in an attempt to follow up with rejoinders. The aim is to produce multiple conversation turns. **Sydorenko, Maynard and Guntly** (2014) suggested that the computer-delivered DCT is superior to the traditional paper-based DCT in assessing pragmatic competence in that the former elicits simulated and extended discourse in a more authentic way.

The importance of simulated tasks can be more salient in occupation-related English language tests or English for special purposes (ESP) tests. One example is an oral communication test in aviation English for air traffic controllers developed by Park (2015). The test simulates a control tower as a virtual assessment environment in Second Life, an online 3D virtual world. In this role-play task, test takers act as air traffic controllers and give oral directives based on incoming aural information. While Park's tasks rely on input that has been recorded, that is, the task is asynchronous, it is feasible to enable multi-user voice communication in a virtual environment like Second Life. In that situation, test takers' interactional competence could also be elicited and assessed through technology-mediated communication.

As can be seen, there are numerous variants of simulated tasks. Some assess all four of the constructs of oral communication more effectively than others. Of particular note is that a major aim of these tasks is to assess interactional competence, but in some cases it is not clear to what extent they can actually be used to measure this ability.

III.4 Integrated tasks

Integrated tasks aim to measure more than one subskill. Examples are listen-speak or read-speak tasks. Developers of these tasks recognize that oral communication rarely involves one-way speech, such as a monologic oral presentation with no question and answer session. These tasks normally include extended written or oral stimuli after which the test taker is expected to provide an extended response. We note that many of the task types that we have discussed require both speaking and listening (which is the major reason we use the term 'oral communication' as opposed to 'speaking' throughout the paper). Integrated tasks can be thought of as an extension of the task type of highly structured speech samples in Harris (1969)'s classification. In this paper, to avoid terminology confusion, we limit the term to the tasks that require test takers to produce speech samples based on given input materials without any synchronous interactions.

Integrated tasks have attracted a great deal of attention from researchers and test developers partly due to the influence of the TOEFL iBT which uses this type of test task to assess speaking ability. In an integrated oral communication task, test takers are required to either listen to a short audio clip or read a short passage, and then summarize the input for a hypothetical audience who does not have access to the same input. Since no interlocutor is needed in the testing process, integrated tasks can be computerized, as exemplified in the TOEFL iBT speaking test. In the integrated tasks of the TOEFL iBT speaking test, computers are used to deliver aural and textual input materials and to record a test taker's speech sample responses. These summary-type tasks have gained some popularity in recent years, in part because of their potential to be rated by automated scoring systems. An example of an automated scoring system is SpeechRater^{SM}, which is currently used to score the speaking section of the TOEFL Practice Online (TPO).

The speech samples elicited from integrated tasks, such as the read-listen-speak task used in the TOEFL iBT, can be rated for phonology, fluency, and grammar and vocabulary. However, this type of task does not directly measure interactional

competence. In addition, using aural or textual input in integrated tasks can complicate the test-taking performance and may make it difficult to determine what the task is measuring. For instance, in such tasks it is not clear to what extent reading comprehension is assessed, and how much working memory capacity affects a test taker's oral performance. The test taker's strategy use in integrated tasks may also be different from other task types (Barkaoui, Brooks, Swain, and Lapkin 2012). The questions about sub-constructs measured with integrated tasks could be more noteworthy when automated scoring tools are used. In the latest version of SpeechRater, the features used for scoring include speech articulation rate, average length of speech chunks, unique words normalized by speech duration, Acoustic Model scores, and Language Model scores. Considering the limitations in construct representation and model prediction accuracy, Xi, Higgins, Zechner, and Williamson (2012) only endorse applications of SpeechRater in low-stakes contexts.

III.5 Elicited imitations

Elicited imitation tasks require a participant to listen to a sentence and then repeat the stimulus material (a word, phrase, or sentence) as closely as possible. This task type was commonly used decades ago but, probably because it is not in line with communicative language teaching principles, fell out of popular use until recently. The revival of elicited imitation tasks in language testing is likely attributable to the ease of delivering and scoring these tasks with automated speech scoring systems. An example of one of these systems is Duolingo's, which is an online language learning website and mobile app. The system uses elicited imitation tasks for its English Test (Ye 2014). These tasks types can be scored using automated speech scoring systems which extract multiple acoustic and prosodic features from test takers' speech samples (Bernstein 2013).

Elicited imitation tasks may provide good estimates of a test taker's fluency and pronunciation through use of automated speech recognition (ASR) technology, but they have limited potential for assessing vocabulary and grammar and little or no potential for assessing interactional competence. While elicited imitation tasks can be reliably scored, and with ASR technology are quite practical, they have been criticized for not having the potential for assessing a broad construct of oral communication (Chun 2006;

O'Sullivan 2013). The task formats and the expected responses in elicited imitation tasks do not involve any interactional aspects of real-life oral communication. Moreover, it can be argued that these task types may be poor indicators of phonology and fluency, since it may be possible for a test taker to simply imitate the phrases with no understanding or ability to segment the speech stream into meaningful parts. In short, imitation tasks, such as sentence repetition tasks, are generally believed to have little potential to assess a broad construct of oral communication. It should also be noted that these tasks could result in negative washback on instruction, since to prepare for such tasks, test takers may spend their time repeating sentences rather than using their time to engage in meaningful discussions with other language users.

IV. CONCLUSION

Assessing oral communication is a rather complicated process, as shown in Figure 1 at the beginning of the paper. A review of the popularly used tasks for assessing oral communication suggests that a number of factors should be considered in determining which task types to include in a speaking test (see Table 1 for a summary of the oral communication testing tasks) for a particular context. Firstly, a clear construct definition should be elaborated. In other words, it is necessary to spell out what should be counted as oral communication in a particular context. We propose that at least four key aspects of oral communication should be assessed, namely, interactional competence, grammar/vocabulary, phonology, and fluency. Secondly, testing tasks and the corresponding scoring rubric should be reviewed with reference to the constructs. The task types listed in Table 1 have been briefly reviewed in this paper and summarized in Table 1. A check mark indicates that this task has good potential for assessing the ability, a question mark indicates that it has limited potential to assess the ability or potential to assess only certain aspects of the ability, and an X indicates that the task has little or no potential to assess the ability. Since each task type has its own merits and drawbacks, our general suggestion is that, after considering the tasks that might be most appropriate for a particular context based on the extent to which they assess all aspects of the construct and their feasibility, test developers should use more than one task type to best ensure construct representativeness.

Table 1. Summary of the characteristics of oral communication testing tasks

Tasks	Example of task format	Constructs measured	
Oral proficiency interviews	face-to-face interviews or phone interview with an examiner in ACTFL OPI	Interactional competence	?
		Fluency	☑
		Grammar & vocabulary	☑
		Phonology	☑
Paired or group oral discussions	unstructured discussion among peers in CET-SET	Interactional competence	☑
		Fluency	☑
		Grammar & vocabulary	☑
		Phonology	☑
Simulated tasks	mini-lecture presentation and question answering in the TEACH test for international teaching assistants	Interactional competence	?
		Fluency	☑
		Grammar & vocabulary	☑
		Phonology	☑
Integrated tasks	summarization after listening to or reading input materials in TOEFL iBT Speaking test	Interactional competence	?
		Fluency	☑
		Grammar & vocabulary	☑
		Phonology	☑
Elicited imitations	sentence repetition in Duolingo English Test	Interactional competence	☒
		Fluency	☑
		Grammar & vocabulary	?
		Phonology	?

REFERENCES

Abeywickrama, P. 2013. "Why not non-native varieties of English as listening comprehension test input?" *RELC Journal 44*(1), 59-74.

ACTFL. 2012. *ACTFL Proficiency Guidelines 2012 – Speaking*. Retrieved from http://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012/english/speaking

Alastuey, M. C. B. 2011. "Perceived benefits and drawbacks of synchronous voice-based computer-mediated communication in the foreign language classroom". *Computer Assisted Language Learning*, *24*(5), 419–432.

Bachman, L. F. 2001. *Speaking as a realization of communicative competence*. Paper presented at the meeting of the American Association of Applied Linguistics – International Language Testing Association (AAAL-ILTA) Symposium. St. Louis, Missouri.

Barkaoui, K., Brooks, L., Swain, M. and Lapkin, S. 2012. "Test-takers' strategic behaviors in independent and integrated speaking tasks". *Applied Linguistics*, *34*(3), 304–324.

Batty, A. 2006. "An analysis of the relationships between vocabulary learning strategies, A Word Associates Test, and the KEPT". *Studies in Linguistics and Language Education: Research Institute of Language Studies and Language Education, 17*, 1-22.

Bernstein, J. C. 2013. "Computer scoring of spoken responses". In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Wiley-Blackwell.

Bonk, W. and Ockey, G. J. 2003. "A many-facet Rasch analysis of the second language group oral discussion task". *Language Testing*, *20*(1), 89–110.

Bosker, H. R., Pinget, A.-F., Quene, H., Sanders, T. and de Jong, N. H. 2012. "What makes speech sound fluent? The contributions of pauses, speed and repairs". *Language Testing*, *30*(2), 159–175.

Brooks, L. 2009. "Interacting in pairs in a test of oral proficiency: Co-constructing a better performance". *Language Testing*, *26*(3), 341–366.

Brooks, L. and Swain, M. 2015. "Students' voices: The challenge of measuring speaking for academic contexts". In B. Spolsky, O. Inbar, and M. Tannenbaum (Eds.), *Challenges for language education and policy: Making space for people* (pp. 65–80). New York: Routledge.

Brown, A. 2003. "Interviewer variation and the co-construction of speaking proficiency". *Language Testing*, *20*(1), 1–25.

Carey, M. D., Mannel, R. H. and Dunn, P. K. 2011. "Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews?" *Language Testing*, *28*(2), 201–219.

Chun, C. W. 2006. "Commentary: An analysis of a language test for employment: The authenticity of the PhonePass Test". *Language Assessment Quarterly*, *3*(3), 295–306.

Elder, C. and **L. Harding.** 2008. "Language testing and English and an international language: Constraints and contributions" in Sharifian, F. and M. Clyne (eds.): *Australian Review of Applied Linguistics (special forum issue) 31*(3), 34.1–34.11.

Fulcher, G. 2003. *Testing second language speaking*. London and New York: Longman.

Ginther, A., Dimova, S. and Yang, R. 2010. "Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring". *Language Testing*, *27*(3), 379–399.

Harris, D. P. 1969. *Testing English as a Second Language*. New York: McGraw-Hill Book Company.

He, L. and Dai, Y. 2006. "A corpus-based investigation into the validity of the CET-SET group discussion". *Language Testing*, *23*(3), 370–401.

Hunston, S., Francis, G. and Manning, E. 1997. "Grammar and vocabulary: Showing the connections". *ELT Journal, 51*(3), 208-216.

Jamieson, J., Eignor, D. R., Grabe, W. and Kunnan, A. J. 2008. "Frameworks for a new TOEFL". In C. A. Chapelle, J. Jamieson, and M. K. Enright (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 55–95). New York: Routledge.

Johnson, M. and Tyler, A. 1998. "Re-analyzing the OPI: How much does it look like natural conversation?". In R. F. Young and A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 27–52). Amsterdam and New York: Benjamins.

Kenyon, D. M. 1992. *Development and use of rating scales in language testing.* Paper presented at the annual meeting of the Language Testing Research Colloquium. Vancouver, Canada.

Leaper, D. A. and Riazi, M. 2014. "The influence of prompt on group oral tests". *Language Testing*, *31*(2), 177–204.

Lin, H. 2014. "Computer-mediated communication (CMC) in L2 oral proficiency development: A meta-analysis". *ReCALL*, 1–27.

Liskin-Gasparro, J. E. 2003. "The ACTFL Proficiency Guidelines and the Oral Proficiency Interview: A brief history and analysis of their survival". *Foreign Language Annals*, *36*(4), 483–490.

McNamara, T. 1996. *Measuring second language performance*. London: Longman.

Nakatsuhara, F. 2011. "Effects of test taker characteristics and the number of participants in group oral tests". *Language Testing*, *28*(4), 483–508.

Nation, I.S.P. 1990. *Teaching and learning vocabulary*. Boston, MA: Heinle & Heinle.

O'Sullivan, B. 2002. "Learner acquaintanceship and oral proficiency test pair-task performance". *Language Testing 19*(3), 277-295.

O'Sullivan, B. 2013. "Assessing speaking". In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 156–171). Wiley.

Ockey, G. J. 2009. The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing*, *26*(2), 161–186.

Ockey, G. J., & French, R. 2014. From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics*. doi: 10.1093/applin/amu060.

Ockey, G. J., Koyama, D., Setoguchi, E., and Sun, A. 2015. The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing*, *32*(1), 39–62.

Ockey, G. J., Koyama, D., & Setoguchi, E. 2013. Stakeholder input and test design: A case study on changing the interlocutor familiarity facet of the group oral discussion test. *Language Assessment Quarterly*, *10*(1), 1–17.

Papajohn, D. 1999. The effect of topic variation in performance testing: The case of the chemistry TEACH test for international teaching assistants. *Language Testing*, *16*(1), 52–81.

Park, M. 2015. *Task selection and simulation for authentic aviation English assessment*. Paper to be presented at the invited colloquium "Task-based language assessment in practice: Justification and realizations" at the 6th International Conference on Task-Based Language Teaching. Leuven, Belgium. September 16-18, 2015

Plakans, B. S. and Abraham, R. G. 1990. The testing and evaluation of international teaching assistants. In D. Douglas (Ed.), *English language testing in U.S. colleges and universities* (pp. 68–81). Washington, DC: NAFSA.

Sato, M. 2014. Exploring the construct of interactional oral fluency: Second Language Acquisition and Language Testing approaches. *System*, *45*, 79–91.

Skehan, P. (1998). Processing perspectives to second language development, instruction, performance, and assessment. *Thames Valley Working papers in Applied Linguistics, 4*, 70-88.

Smith, L. and Bisazza, J. 1982. The comprehensibility of three varieties of English for college students in seven countries. *Language Learning, 32(2),* 259-269. **Swender, E.** 2003. Oral Proficiency Testing in the Real World: Answers to Frequently Asked Questions. *Foreign Language Annals*, *36*(4), 520–526.

Sydorenko, T., Maynard, C. and Guntly, E. 2014. Rater behaviour when judging language learners' pragmatic appropriateness in extended discourse. *TESL Canada Journal*, *32*(1), 19–41.

Turner, C. 2009. Examining washback in second language education contexts: A high stakes provincial exam and the teacher factor in classroom practice in Quebec secondary schools. *International Journal of Pedagogies and Learning, 5*, 103-123.

Van Lier, L. 1989. Reeling, writhing, drawling, stretching, and fainting in coils: oral proficiency interviews as conversation. *TESOL Quarterly*, *23*, 489–508.

Winke, P. 2013. The effectiveness of interactive group orals for placement testing. In K. McDonough and A. Mackey (Eds.), *Second Language Interaction in Diverse Educational Contexts* (pp. 247–268). Amsterdam/Philadelphia: John Benjamin Publishing Co.

Wu, S.-L. and Ortega, L. 2013. Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals*, *46*(4), 680–704.

Xi, X., Higgins, D., Zechner, K. and Williamson, D. 2012. A comparison of two scoring methods for an automated speech scoring system. *Language Testing*, *29*(3), 371–394.

Ye, F. 2014. *Validity, reliability, and concordance of the Duolingo English Test*. Pittsburgh, PA. Retrieved from https://s3.amazonaws.com/duolingo-certifications-data/CorrelationStudy.pdf

Zheng, Y. and Cheng, L. 2008. Test review: College English Test (CET) in China. *Language Testing*, *25*, 408–417.

Received: 3 June 2015

Accepted: 5 August 2015

Cite this article as:

Ockey, G.J. & Zhi, L. 2015. "New and not so new methods for assessing oral communication". *Language Value* 7 (1) 1-21. Jaume I University ePress: Castelló, Spain. http://www.e-revistes.uji.es/languagevalue. DOI: http://dx.doi.org/10.6035/LanguageV.2015.7.2

ISSN 1989-7103